Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

DISCRIMINATION BETWEEN TWO LUNG DISEASES USING CHEST RADIOGRAPHS

Norliza Mohd. Noor[†], Omar Mohd. Rijal*, Hamidah Shaban[§], Ong Ee Ling*
norliza@citycampus.utm.my , omarrija@um.edu.my,
hamidah_shaban@yahoo.com.uk e_lyn9823@yahoo.com
* Institute of Mathematical Science, University Malaya
[†]Diploma Program Studies, Universiti Teknologi Malaysia
[§]Institute of Respiratory Disease, Kuala Lumpur Hospital

ABSTRACT

Economic considerations make the conventional chest radiograph (X-ray) film an important ingredient in the diagnostic process. An initial clinical investigation for patients with suspected lung ailments is the study of the chest X-rays. The problem of detection for diseases in their early stages are well known using X-ray. A technique involving wavelets coefficient as the feature vector and Andrew's Curve has been proposed for detection of Mycobacterium Tuberculosis (MTB). This paper presents new and important results whereby lung cancer (LC) may be detected and differentiated from MTB. A method to calculate misclassification probabilities is given.

INTRODUCTION

A well known problem [1] with the detection of lung cancer (LC) using conventional chest X-ray is that detection at the early stages is difficult, and a cancer-patient may well be detected in his critical or later stage with fatal consequences. As a first step to reduce the problem, there is a need to detect LC more efficiently using chest X-rays. The efficiency of detection clearly depends on the ability of the medical practitioner to discriminate LC with other lung diseases, in particular Mycobacterium Tuberculosis (MTB).

The authors in [2] looked at subsets of the digitized chest X-ray image of a confirmed MTB patient. In particular, the infected region seen visually as white cloud is the region of interest in this study. Figure 1 shows an X-ray of confirmed lung cancer patient. The cancer is indicated by the white spot located at the middle part of the left lung. Figure 2 shows an X-ray of confirmed MTB patient. The MTB is indicated by white cloudy spot on the upper right lung.

Samples of the infected area are in the form of vertical lines defined between a given pair of adjacent ribs, which in turn is defined as the line profile. The line profiles are obtained manually under advice of medical expert. Thirty line profiles were obtained. For each line profile, one-dimensional discrete wavelet [3] was applied giving the corresponding approximate and detailed Daubechies Coefficients. In total, a vector of 26 coefficients represented each line profile.

A result from [2] suggest the average line profiles, namely

$$\underline{\bar{x}} = \left(\underline{x}_1 + \underline{x}_2 + \cdots + \underline{x}_{30} \Big/ 30 \right)$$

would be used as the feature vector for MTB where $\underline{x}_i$ is the i[th] vector of Daubechies wavelet approximate coefficients. In this study, similar ideas are applied to LC. Only confirmed cases of LC, MTB and normal (healthy) patients are used to provide the control feature vectors. The control feature vector is then presented graphically using Andrew's Curve.

ANDREW'S CURVES

The $\underline{\bar{x}}$-feature vectors may be used as a feature to identify LC and MTB. The **visual** comparison of two 26-dimensional vectors is clearly not appealing. Andrews [4] proposed a method of plotting a data point $\underline{x}_r^T = (x_{r1}, \ldots, x_{rp}), r = 1, \ldots, n$, which involves plotting the curve $\{t, f_{\underline{x}_r}(t)\}$ where

$$f_{\underline{x}_r}(t) = x_{r1}/\sqrt{2} + x_{r2}\sin t + x_{r3}\cos t + x_{r4}\sin 2t + x_{r5}\cos 2t + \cdots$$

for each "data point' $\underline{x}_r (r = 1, \ldots, n)$ over the interval $-\pi < t < \pi$. Thus, each data point (in this case our $\underline{\bar{x}}$-vectors) will appear as a harmonic curve drawn in two dimensions. It may be shown that $\int_{-\pi}^{\pi} [f_{\underline{x}}(t) - f_{\underline{y}}(t)]^2 \, dt$

0-7803-8740-6/05/$20.00 ©2005 IEEE.

between two curves $\{t, f_{\underline{x}}(t)\}$ and $\{t, f_{\underline{y}}(t)\}$ is proportional to the square of Euclidean distance between $\underline{x}$ and $\underline{y}$.

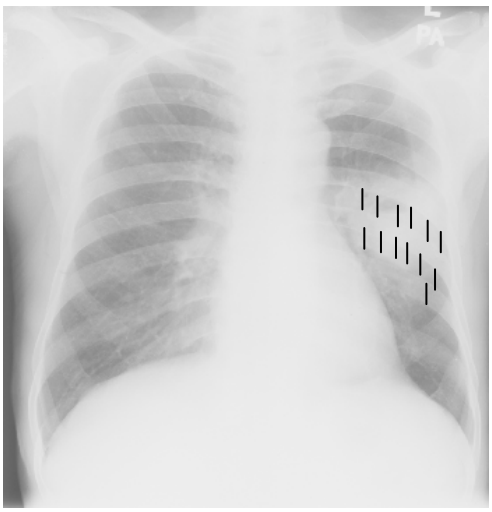

Figure 1 Confirmed lung cancer (left lung) chest X-ray.

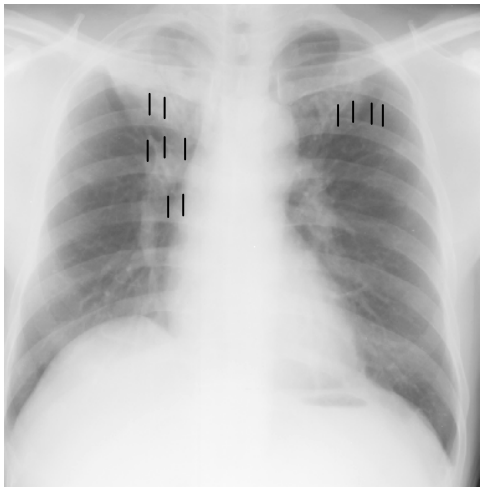

Figure 2 Confirmed MTB chest X-ray.

CLUSTERING WITH ANDREW'S CURVES

When a pair of Andrew's Curve such as $\{t, f_{\underline{x}}(t)\}$ and $\{t, f_{\underline{y}}(t)\}$ for $-\pi < t < \pi$ are close and possibly overlapping we have the case where $\underline{x}$ and $\underline{y}$ have very similar values. This result provides a clustering method.

Figure 3 and Figure 4 shows three groups of curves. The LC infected areas, selected from Figure 1 show denser white cloud compared to that of the MTB infected areas, see Figure 2. This may explain why the red LC curve have largest positive values, for example, at t = 0.1, and smallest negative values at t = -0.25.

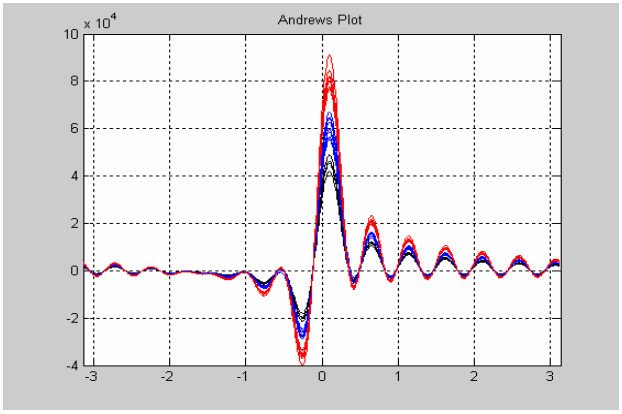

Figure 3 Andrews curve for LC (red line), MTB (blue line) and normal (black line) in the range of $\pi < t < -\pi$.

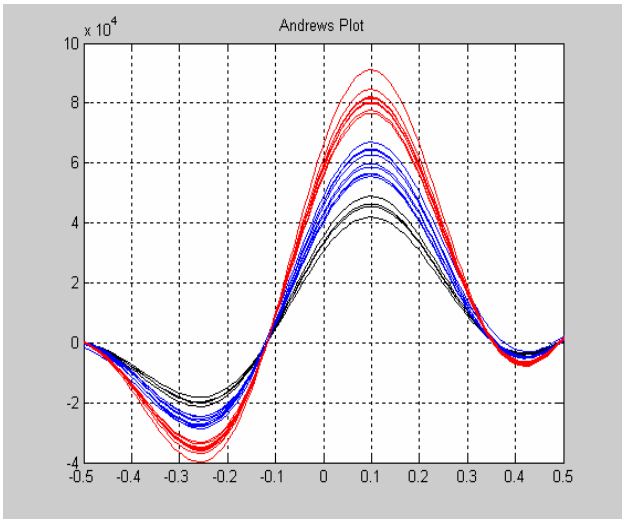

Figure 4 Andrews curve for LC (red line), MTB (blue line) and normal (black line) in the range of $-0.5 < t < 0.5$.

DISCRIMINATION WITH ANDREW'S CURVES

For a given or fixed t-value, the difference between $f_{\underline{x}}(t)$ and $f_{\underline{y}}(t)$ is only due to the difference in the $\underline{x}$ and $\underline{y}$ vectors. For certain t-points the $f(t)$ values have largest variance. For example, when t = 0.1 in Figure 2 each colour group shows good separation with large variance (with respect to each group). These remarks forms the foundation of a discrimination technique

whereby an unknown observation, say $\underline{w}$, may be assigned to one of the colour groups.

For a fixed t-value (e.g. $t = t_0$) let

$$[f_{\underline{x}}{}^{1}(t_0), f_{\underline{x}}{}^{2}(t_0), \cdots, f_{\underline{x}}{}^{k}(t_0)],$$

$$[f_{\underline{y}}{}^{1}(t_0), f_{\underline{y}}{}^{2}(t_0), \cdots, f_{\underline{y}}{}^{l}(t_0)]$$

and $$[f_{\underline{z}}{}^{1}(t_0), f_{\underline{z}}{}^{2}(t_0), \cdots, f_{\underline{z}}{}^{m}(t_0)]$$

be the $f(t_0)$ values for k -LC patients, l -MTB patients, and m –normal patients. Further let m_1, m_2 and m_3 be the median of the $f(t_0)$-values for the LC-group, MTB-group and normal-group, respectively.

Suppose $\underline{w}$ is an unidentified patient. Let $f_{\underline{w}}(t_0)$ be the $f(t_0)$ value for $\underline{w}$. If $f_{\underline{w}}(t_0)$ is closest to

$$\begin{cases} m_1, \underline{w}\ is\ an\ LC-patient \\ m_2, \underline{w}\ is\ an\ MTB-patient \\ m_3, \underline{w}\ is\ an\ Normal-patient \end{cases}$$

Six t-values were studied (see next section) and $f_{\underline{w}}(t_j),\ j = 0,1,\cdots,5$ was seen to be consistently closest to the same median value. This in fact is our assignment rule.

THE CHOICE OF *t*-VALUES FOR DISCRIMINATION

Starting at t = 0, with increment of 0.01 units each time, we calculate for each t –value;

$$D(t) = \min(f_{\underline{x}}(t)) - \max(f_{\underline{y}}(t))$$

for every t –interval of length 0.5. We find t_0 such that $D(t_0) = \max_t D(t)$. Six t –values were obtained.

MISCLASSIFICATION PROBABILITIES

Let $\{\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_k\}, \{\underline{y}_1, \underline{y}_2, \cdots, \underline{y}_l\}$ and $\{\underline{z}_1, \underline{z}_2, \cdots, \underline{z}_m\}$ represent three samples corresponding to LC-patients, MTB-patients and normal patients respectively.

One observation, say $\underline{x}_i\ (1 \le i \le k)$ is removed from the LC-group. The median m_1 is recalculated based on the remaining k-1 vectors. The vector $\underline{x}_i$ is now treated as the unknown observation and assigned to one of the three groups as described in the previous section; and the event of misclassification is noted. This process is repeated for another $\underline{x}$-vector and finally gives the estimated misclassification probability for the $\underline{x}$-vector group.

Misclassification probabilities may also be calculated for the $\underline{y}$-vector group, and then for the $\underline{z}$-vector group. In fact, this study looks at the overall misclassification probability, i.e. number of $\underline{x}$-vector, $\underline{y}$-vector, $\underline{z}$-vector misclassified divided by $k + l + m$.

RESULTS AND DISCUSSION

A novel detection method based on statistical methods and wavelets were applied for the problem of discrimination between two lung diseases using chest radiograph. The motivation of using Andrews Curve to graphically represent the average line profile as a feature to detect MTB is to avoid the difficult task of defining a new MTB feature according to physical characteristic like shape, size and area since the white cloud or snowflakes do not have fixed (or consistently varying) dimensions for such characteristics.

Important decisions frequently have to be made by visually interpreting chest radiographs. This paper has shown that the Andrew's Curve provide an automatic comparison between MTB, LC and normal lung which eliminates problems associated with visual interpretations of chest radiographs. Further, the method suggested in this paper gives misclassification probabilities to indicate the qualities of the method.

In this study 9 LC patients, 10 MTB patients and 4 normal lung patients were studied. Table 1 gives the median m_1, m_2 and m_3 for six t –values corresponding to maximum *D(t)* values. This table may then be used to detect possible LC and MTB cases.

Although robust properties of the above discriminant technique should be further studied, achieving 100% correct classification in our experiment suggests our methods are reliable.

REFERENCES

[1 Shaban, H, Medical Consultant for Respiratory Disease, Private Conversation, Institute of Respiratory Disease, Kuala Lumpur Hospital.

[2] Noor, N. M., Rijal, O.M and Chang, Y.F., "Wavelet as features for Tuberculosis (MTB) using standard x-ray film images", IEEE proceedings of 6th International Conference on Signal Processing (ICSP02), Beijing, China, 2002.

[3] Daubechies, I., Ten Lectures on Wavelets, Society For Industrial and Applied Mathematics, Philadelphia, 1992.

[4] Andrews, D.F., "Plots of high dimensional data", Biometrics, 28, pp.125-36, 1972.

Table 1: Group median values at selected *t*.

***t* - value**	**0.10**	**0.66**	**1.16**	**1.64**	**2.12**	**2.60**
normal	46133	12096	7425	5435.9	4169.4	3257.5
MTB	59076	15148	9436.8	6962.9	5385.1	4170
LC	81385	20689	12888	9538.2	7399.1	5762.2